NASA TECHNICAL NOTE

NASA TN D-5472

C. 1

NASA TN D-5472

COMMUNICATION SYSTEM DESIGN

by Louis A. Frasco

Electronics Research Center

Cambridge, Mass.

NATIONAL AERONAUTICS AND SPACE ADMINISTRATION • WASHINGTON, D. C. • OCTOBER 1969

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
NASA TN D-5472		

4. Title and Subtitle	5. Report Date
Communication System Design	October 1969
	6. Performing Organization Code

7. Author(s)	8. Performing Organization Report No.
Louis A. Frasco	C-83

9. Performing Organization Name and Address	10. Work Unit No.
Electronics Research Center Cambridge, Massachusetts, 02139	125-21-01-02-25
	11. Contract or Grant No.
	13. Type of Report and Period Covered
12. Sponsoring Agency Name and Address National Aeronautics and Space Administration	Technical Note
	14. Sponsoring Agency Code

15. Supplementary Notes

16. Abstract

In this report, a general system design philosophy palatable to both analog and digital communication engineers is presented. The method of presentation is novel and some new results and calculations are presented. An attempt is made to establish the relevance of the more theoretical aspects of communication system design (i.e., statistical communication theory, information theory and coding, etc.) to the work of the practicing system design engineer in evaluating real systems. Consideration of the fidelity of the overall system — from source characterization to user requirements — is of prime importance. The use of the computer in system management of hybrid analog/digital systems is also discussed. In addition, optimum bounds on performance are calculated for the space channel and compared with the performance of some existing systems.

17. Key Words	18. Distribution Statement
. System Design Philosophy . Statistical Communication Theory . Information Theory and Coding . System Management	Unclassified - Unlimited

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price*
Unclassified	Unclassified	31	$3.00

*For sale by the Clearinghouse for Federal Scientific and Technical Information
Springfield, Virginia 22151

COMMUNICATION SYSTEM DESIGN

By Louis A. Frasco
Electronics Research Center

SUMMARY

Many communication system designers feel that they have been short-changed by the communication theorist, asserting that he has not stated his results in terms meaningful to the practicing engineer. The system designer is interested in the overall system performance from the data source to the user. He requires measures of performance expressed in terms meaningful to him; e.g., achievable data rate and error rate. In a given design situation, he may or may not have the freedom to choose all parts of the system. In general, however, he is seeking to optimize the overall system over a class of allowable processors and modems. In this report, an attempt is made to establish the relevance of the theoretical aspects of communication system design to the work of the practicing engineer in the design of real systems. Both analog and digital systems are discussed and lead naturally to the consideration of hybrid systems as a general solution to system design. Also, the use that can be made of the digital computer for data management and executive control, and also in the realization of communication systems is pointed out. The report is mainly tutorial with particular emphasis on clarity and ease of understanding throughout.

INTRODUCTION

In this report, a general introduction to the reliable communication of information through noisy media -- space communication channels, in particular -- is presented. Both random corruption and quantization errors due to discrete representation of continuous functions are considered. A variety of measures of reliability are looked at and their relative merits discussed realistically with regard to both system constraints and user requirements. The selection of suitable fidelity criteria in communication system design usually involves a high degree of subjectivity and is generally a difficult problem. Its fundamental importance, however, requires that it be carefully considered.

An attempt is made to establish the relevance of the more theoretical aspects of communication system design (i.e., statistical communication theory, information theory and coding, etc.) to the work of the practicing system design engineer in evaluating real systems.

Optimum bounds are calculated for the space channel with 2-level and ∞-level quantization at the receiver using information bit error probability P_b as a measure of reliability. The results yield an interpretation of the trade-offs between P_b, SNR, and information rate. In addition, these bounds are plotted and a comparison made with linear codes (2-level), convolutional codes with sequential decoding (∞-level), and orthogonal/biorthogonal codes with matched filter reception.

SOURCE CHARACTERIZATION

An information source (e.g., the outputs of which are meter readings, particle counts, speech or video waveforms, etc.) is to be interpreted at some remote location. The fidelity required for the overall system in reproducing this source depends on the requirements of the user. For example, consider transmitting the following English text: M = CAN YOU HEAR ME? which gets garbled by the transmission medium to $\hat{M}$ = CEN YOX HE$R MAY?. If it is only required that the receiver be able to understand the transmitter and exact reproduction is not demanded, some errors can be allowed and M can still be reconstructed (decrypted) from $\hat{M}$ due to the redundancy of the English language. Thus, the degree of reliability required in transmitting a source is intimately related to its information content; i.e., that part of the message which is not redundant. Whatever portion of the message can be determined from this part is not worth sending. This concept of information content can be formalized and is the basis for the theoretical constraints on all so-called data-compression schemes. This is just another name for the general source coding problem which shall be discussed a little later.

The transmission of scientific measurement data from remote locations is another example. In a particle counting experiment, should a scientist be concerned if an error of a percent or so is made in the received particle count? The answer depends on the experiment. He may be trying to calculate a fundamental constant or predicting a third-order effect in some process. He might, however, only be looking for large changes in particle count in order to detect some gross physical effect. The particular case makes a great deal of difference in how reliability is measured. In the context of everyday life, this amounts to the differences in the tolerances a person might allow in measuring out a few feet of rope and those found in a precision watch. This is an essential part of numerical approximation. Everyone is aware of the importance of being able to make quick approximations which are accurate within the context of their daily lives.

Some additional comments on redundancy fundamental to the information transmission problem should be mentioned. In the first example, where English text was transmitted, it was seen

that the text can still be reproduced after an appreciable amount of corruption. However, an extremely complicated system was used to accomplish this -- the human mind. Therefore, while not requiring exact reproduction, it is found that the text can still be recovered, but by using an extremely complex deciphering process. In general, it will be found that an increase in system complexity or computation will have to be traded for a corresponding increase in source information rate.

Fundamental to the communications problem is source characterization. What properties of the source are relevant to the observer? As has been mentioned, this question depends on the particular problem at hand. In a particle-counting experiment on board a spacecraft the observer-scientist might be interested only in large fluctuations in particle count. Similarly in image processing, the observer is probably very interested in changes in intensity which correspond to the boundaries of regions in the image. Once the properties of the source which are of interest in a particular context have been determined, an attempt is made to describe its output at the receiver within the tolerances imposed by the user. In a fuzzy sort of way, it should be clear that an attempt to characterize the source based only on these properties will, in general, be easier than requiring complete specification. This sort of fuzziness will be tolerated; it can be eliminated but the effort adds nothing to one's understanding.

At this point a stochastic model must be adopted to describe the uncertainty associated with an inability to adequately describe the source output at a given time. For if there were no undertainty and the source output could be specified at each instant of time within user tolerances, there would be no information to transmit! Certainly one should start by trying to fit observations to the simplest models. A few questions come readily to mind. Are the source statistics independent of the time of observation (stationarity)? Is their distribution uniform, Gaussian, Poisson, etc.? If a stationary source is assumed (which is realistic in many situations), what does the source power spectrum look like? Is it band-limited? The answers to these questions are fundamental to the design of an efficient communication system.

The statistical distribution of the source output may be determined basically in two ways: (1) A theoretical analysis of the underlying physical processes producing the source outputs; for example, one can model the particle-counting experiments as measurements from a source the output of which is a Poisson process; (2) The experimental measurement of source outputs used to determine an empirical distribution. In this case, output data can be used to construct histograms or used to estimate parameters of an assumed distribution.

Well-known techniques have been developed to measure the power spectrum from source outputs. The shape of the source spectrum is important for the following reasons. If the source output is band-limited (i.e., its power spectrum is limited to frequencies within a finite band), it can be shown that no information is lost if the source output is sampled at an appropriate rate (Nyquist rate). Moreover, if the power spectrum is flat, the samples will be statistically uncorrelated (which for a Gaussian source implies statistically independent samples). But what does all of this mean in terms of communicating the source to the user?

The output of a general source is an analog waveform. At any fixed instant of time, this waveform is theoretically capable of carrying as much information as one desires. For example, if the value the waveform takes on at a particular instant could be measured exactly, this complete report could be recorded as a number equal to this value by the following simple construction. Give a binary code to each letter including the colon, the period, the space, etc. (as a digital computer does). Then, starting with the first word in the report and ending with the last, string out the corresponding sequence of binary digits, put a period in front, and consider it a binary fraction the value of which is the height of the waveform at the fixed instant! Of course, this proposal is unrealistic but it does bring home the following point. The problem is not with the construction, but it is inherent in an inability to make measurements accurate enough. This appears to be a separate physical problem from the noise, but is nevertheless of the same general type. Also, in reality, the source is measured with real instruments of finite bandwidth which constrains the amount of "wigglyness" they can interpret. This means that already a trade-off exists between instrument bandwidth and the bandwidth necessary to describe the source. In any case, in the real world, all signals are band-limited and, therefore, there is some upper limit to the rate at which they must be sampled without any loss excluding the loss due to the measuring instrument.

Finally, the added niceties of a flat source spectrum and Gaussian statistics allow the source to be faithfully represented by a sequence of independent Gaussian time samples. Since adjacent samples contain no information about each other (no coupling) due to their statistical independence, the original analog source has been simply represented as a memoryless time-discrete source by time sampling at the Nyquist rate. In general, for an arbitrary source spectrum the time samples will be correlated. However, as shall be seen later, it may be profitable to do some preprocessing of the source to try to achieve some of those qualities which seem desirable (e.g., prewhitening, distribution transformation, etc.) It should be clear at this

point that source characterization is fundamental and requires some care; sloppiness at this point will propagate added problems throughout the whole system. A by-product of source characterization is the implicit determination of realistic reliability criteria. Next some representative performance criteria are considered.

MEASURE OF DISTORTION

The amount of source distortion that can be tolerated is entirely dependent on user requirements. These requirements aid us in our choice of a distortion measure. Consider a general source with output a(t), a sample function from a stochastic process. The source output is somehow processed, transmitted, reprocessed, and then presented to the user as $\hat{a}(t)$ (see Figure 1). The instantaneous error between the actual source output and its reproduction at the receiver is defined by $e(t) \triangleq a(t) - \hat{a}(t)$; e(t) is then the deviation from exact reproduction. The problem now reduces to choosing a distortion function D[e(t)] which measures the user's relative happiness with what he has received. This choice may be difficult. Choosing a specific distortion function which meets user requirements is somewhat subjective. Can there be any assurance that exactly the right measure of distortion has been chosen? Probably not. At best there exists a fuzzy relationship between user requirements and the proper distortion function. It can only be hoped at the outset that system performance is not particularly sensitive to the detailed form of the distortion function chosen -- only to its general functional form. One can state two general properties "nice" distortion functions might have: 1) They should be monotone-increasing; i.e., the larger the error e(t) the larger the distortion; 2) They should also be symmetric; i.e., positive and negative errors of the same magnitude yield the same distortion. Even if these properties do not seem reasonable, in general, there exists a large class of distortion functions which satisfy them. Some examples of distortion functions are shown in Figure 2. Since a(t) is a sample function from a random process, the statistical average of the distortion measure $\overline{D[e(t)]} \triangleq E[D[e(t)]]$ shall be considered. This does not impose any serious restriction on the fidelity criteria that can be considered. For example: a) $\overline{D_1[e(t)]} = E[|a(t)-a(t)|]$, the mean error magnitude; b) $\overline{D_2[e(t)]} = E[(\hat{a}(t)-\hat{a}(t))2]$, the mean square error; c) $\overline{D_3[e(t)]} = A\ Pr[|a(t)-\hat{a}(t)|>d]$, the probability that the error magnitude is greater than some constant (for the case A=1). In the particular case when the source statistics and channel disturbance are Gaussian, it has been shown that the form of the optimum processor (receiver) is invariant over a large class of distortion functions (including those of Figure 2). This is particularly satisfying, since whatever distortion measure was subjectively chosen from this large class, it would lead to an identical system.

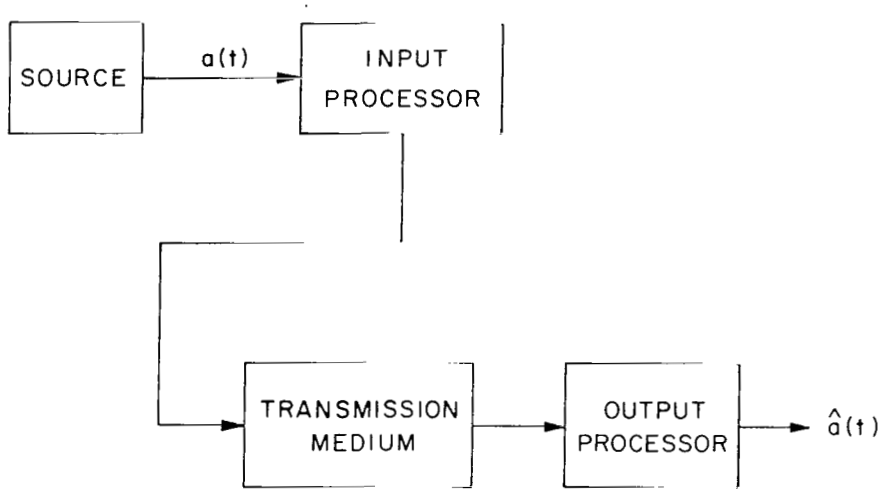

Figure 1.- General communication system model

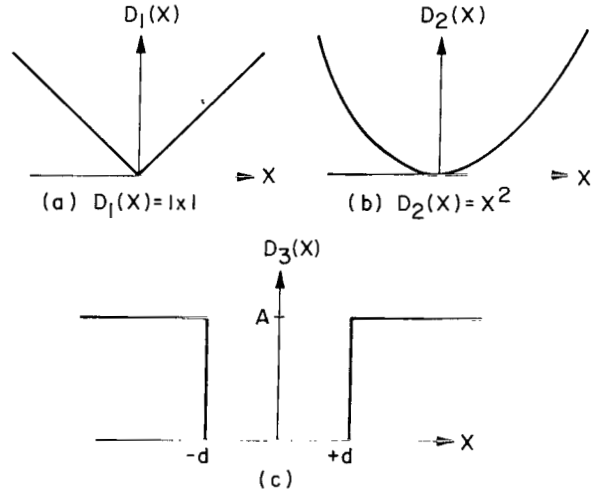

Figure 2.- Distortion functions

The previous examples of distortion measure have all been passive. In general, there may be interest in distortion measures which include the dynamics of the error. For example, in the particle-counting experiment where rapid fluctuations in the count are of interest or in image processing where the intensity variation at boundary regions in the image is of particular importance, one might want a distortion function of the form:

$$D[e(t)] = |e(t)| + \alpha \frac{d\,|e(t)|}{dt} .$$

The number of different types of distortion functions is large and diverse. The importance of the proper choice of distortion function (or at least choice of a class of functions to which performance is relatively insensitive) cannot be over-

emphasized. The added effort required to pick a distortion function which accurately characterizes the source to within user requirements is repaid many times over in the overall system design function.

The problem of reliable communication has now been reduced to constraining the average distortion $\overline{D[e(t)]}$ associated with the information source to a tolerable level. A reasonable question to ask is how well can one hope to do; i.e., what is the optimum performance. This is a very important question and should be answered. The answer will yield a yardstick by which to judge the performance of sub-optimum systems.

RATE-DISTORTION THEORY

Questions relating to optimum source reproduction for a given distortion measure find their answers in Shannon's rate-distortion theory (an area of information theory which, unfortunately, is practically unknown -- and hence obscure -- to anyone other than information theorists). In the next few lines, a cursory treatment of rate-distortion theory relevant to the problem at hand will be attempted.

Once a distortion measure $D[e(t)]$ is chosen for the source and a tolerable level of distortion ε (i.e., $\overline{D[e(t)]} < \varepsilon$) determined, a function $R(\varepsilon)$ can be defined which depends on the source statistics, distortion measure, and, of course, ε. $R(\varepsilon)$ is called the *rate-distortion function*. It can be shown that $R(\varepsilon)$ measures the equivalent information rate of the source (e.g., information bits/sec). If everything between the source and the user is considered as the channel (i.e., processors and transmitting medium), it can be characterized by a transition probability distribution between source outputs and the inputs to the user, and a quantity C called *channel capacity* can be defined. Channel capacity is a quantity fundamental to "conventional" information theory and measures the maximum rate at which information may be transmitted over the channel reliably (i.e., in the sense of arbitrarily small error probability). It can be shown that a necessary condition for transmitting a waveform a(t) over a channel with capacity C with an average distortion ε or less is that $R(\varepsilon) \leq C$ (result due to Shannon). In fact, it is possible to encode the output of the source and transmit it over a channel with capacity C with a distortion as near ε as desired for any $R(\varepsilon) \leq C$! The above theorem is what gives meaning to $R(\varepsilon)$ as the *equivalent rate of the source*, i.e., the highest rate at which the source must be transmitted and still keep the average distortion ε or less.

These results are very satisfying. Only the calculation of $R(\varepsilon)$ for a given source and distortion measure is needed in order to determine the optimum (maximum) information rate of the source for a given distortion ε. There is a catch, however. In general,

$R(\varepsilon)$ is very difficult to calculate for an arbitrary source. Presently much of the work being done in rate-distortion theory is in determining methods to calculate $R(\varepsilon)$ or to find tight bounds. Also, some effort has been devoted to defining $R(\varepsilon)$ for a class of sources in cases where only vague *a priori* knowledge is available about the source. $R(\varepsilon)$ has been calculated in some cases -- for the Gaussian source with mean square error distortion measure in particular. In this case, for a source the outputs of which are band-limited, setting $R(\varepsilon) = C$ where C is the capacity of a band-limited channel with additive white Gaussian noise, an equation is obtained for the minimum mean square error distortion as a function of required signal-to-noise ratio. It can be shown that this result for the Gaussian case upperbounds the optimum performance for all other sources using the mean square error criterion. This result can be compared with the curves for sub-optimum processors to measure their performance. If the source is not Gaussian or the appropriate distortion measure is not in the mean square error class (e.g., Figure 2), *or* both, then the calculation of $R(\varepsilon)$ may be very difficult. Therefore, as a matter of general design philosophy, one should compare, if possible, the performance of each sub-optimum system considered for a particular system application with the best that can possibly be done.

COMMUNICATION SYSTEM CLASSIFICATION

Processors of a general stochastic source are of three basic types: 1) Analog (continuous); 2) Time Discrete - Amplitude Continuous (sampled-data); 3) Time Discrete - Amplitude Discrete (sampled and quantized). Examples of each type are listed in Figure 3. Of course, analog processing is the most general in the sense that one is able to deal with a continuous source output. In general, processing becomes more restrictive (less efficient but probably also less complicated) as one moves from Type 1 to Type 3 processing. However, if the source is band-limited, nothing is lost by using a Type 2 processor sampling at the Nyquist rate. As mentioned previously, in practice one deals with sensors and associated hardware with finite bandwidth and, therefore, must live with this bandwidth constraint. Therefore, without much loss of generality, the processor input stream can be considered to consist of continuous samples of the source output taken at uniform times (these samples have as their probability distribution, the distribution of the source, and therefore are, in general, correlated). In moving to Type 3, one must be more careful. The transition from Type 2 requires that each time-discrete sample (or sequence of such samples) be quantized and then processed. How much information is lost, if any? The answer depends on a lot of things -- the particular source statistics, the channel noise, and the distortion measure. For example, if a Gaussian source and the mean square error distortion criteria are assumed, it has been shown that even for

TYPE	EXAMPLES
1	Optimum Angle Modulation AM-Amplitude Modulation FM-Frequency Modulation
2	PAM-Pulse Amplitude Modulation DFM-Discrete Frequency Modulation PDM-Pulse Duration Modulation
3	PCM-Pulse Code Modulation Algebraic Coding-Hamming and BCH Codes Probabilistic Coding-Sequential Decoding

Figure 3.- Communication system classification

uncorrelated time samples processed by an optimum uniform quantizer, about 1/4 bit per sample more is required than with optimum processing of sequences of the continuous amplitude samples. In general, the type of processor must be chosen based on the particular problem and the performance required.

Up until now, little mention has been made of some of the more practical aspects of communication system design. Along with theoretical performance of optimum processors, important questions of complexity, particular system requirements, compatibility, etc., must be considered. These questions, as one might suspect, are much more difficult to answer in general. Most of the results one gets in information theory concerning optimum processors are based on existence proofs. They are not constructive. One is not given anything which can be built (or at least anything simple enough so that one would want to!).

In summary, the communication system designer is generally confronted with choosing among analog systems (Types 1 and 2) and digital systems (Type 3) in a particular design situation. In the past few pages, an attempt has been made to convey the inherent depth of the problem and to point out that the intelligent designer had best consider the whole problem from source characterization to user requirements. Arguments can surely be made for both analog and digital communication, but it is foolish and very dangerous (in terms of system efficiency) to stick stubbornly to one scheme or the other, regardless of the problem at hand.

HYBRID SYSTEMS

In general, a communication system designer will find himself combining both types of processing into some sort of hybrid scheme which utilizes the advantages of each. For example, analog processing may be "natural" for spectral shaping and various

other aspects of signal conditioning. Some examples of "natural" processing include controlling picture resolution with film of different sensitivity (the film contains the squared magnitude of the source output spectrum), the use of optical filters, and mechanical linkages in measuring instruments (e.g., governors or stops to control dynamic range, etc.). As an example of trade-offs between analog and digital techniques, consider a particular application where eight levels of quantization (i.e., 3-bit quantization) of the source data samples introduce negligible degradation from the continuous source representation (based on overall system performance). In this case, the levels could be efficiently encoded into sequences of binary digits (either by optimum (entropy) source encoding techniques or, suboptimally, simply as the binary number representation of the level). The binary digits could then be grouped into sequences of length k, transmitted over the channel as one of $M=2^k$ analog waveforms and then detected at the receiver, converted back to k binary digits and decoded as a quantization level. Alternatively, one could take the block of k binary digits and send them one at a time by using only 2 analog signals instead of 2^k. This method is commonly called bit-by-bit signalling. Of course, if one of the bits sent is inverted by noise on the channel, the error will not be detected. To get around this, a parity bit can be added on before the block is transmitted to tell if the number of 1's in the block is odd or even. Then the block of k+1 bits is sent bit-by-bit; if a single error occurs, it will be detected. In general, additional bits can be added to each block to obtain more information. Therefore, each sequence of k information bits may be encoded into blocks n bits long (where the additional n-k bits have been added to combat noise). These blocks are called *codewords* and the corresponding code structure, an (n,k) block code of length n with k information bits. The *code rate* (in information bits/channel bit) is defined as R=k/n and is a measure of the redundancy of the block code. It is clear that a trade-off exists between the rate R at which information is transmitted with a block code and the increase in performance due to the additional bits sent over the channel to combat noise. The underlying philosophy in the foregoing example is fundamental. Given output data samples from the source, as much redundancy is eliminated from them as possible (data compression). Then based on the noise statistics of the channel, a certain amount of "structured" redundancy is introduced to fight channel noise (e.g., parity bits).

It must be emphasized again that the preceding examples are just meant to point out the flexibility available in system design. They are not meant to serve as an exhaustive list of the alternatives available and are, in general, suboptimal. What is actually done in a particular situation will depend on the level of reliability demanded and practical considerations. The task of the communication system designer is not an easy one. At best, one

can equip him with an incomplete specification of the problem, a multitude of subjective considerations, an intelligent set of alternatives, courage, and hope for the best!

At this point, an impasse has been reached in the general discussion of analog and digital communication systems. In order to continue, the discussion must be restricted to a particular system type. The choice is arbitrary in that both types of systems deserve consideration in a general system design problem. However, analog communication has reached a reasonable level of sophistication (in theory at least, if not in application) and many good books have been written on the subject both for the theoretician and the practicing engineer. Along with these considerations, the fact that digital communication techniques are less well-known by communication system designers justifies an in-depth study of these techniques. Of course, whenever appropriate, various trade-offs between the system types will be discussed.

DIGITAL COMMUNICATION SYSTEMS

There are many system considerations which make digital communication very attractive. For example, in digital systems where bit-by-bit signalling is used (± some analog waveform) propagation loss in the transmission medium is easily handled by repeaters placed along the transmission path. A repeater is a device which decides whether the attenuated or distorted waveform is positive or negative and then regenerates it. Also, with the ever-increasing role being played by the digital computer today (not to outlaw analog computers where appropriate), information not in digital form to start with is very likely to run into a digital computer somewhere along the line and have to be digitized. In the design of communication systems for space probes the computer is a powerful tool for data management of experiments (e.g., formatting, preprocessing, adaptive processing, reconfiguring, etc.). Of course,one could argue that in certain cases the computer could be used efficiently to control peripheral analog devices. In general and from a more mundane point of view, a digital processor usually turns out to be conceptually simpler to understand. Complicated processes may be simply expressed numerically for interpretation by computer.

In communicating source information to the user, a good deal of the performance depends on the channel statistics as well as source statistics. In dealing with space communications, the only limitation imposed by the channel is essentially the receiver's front-end thermal noise (all other sources of disturbance such as sky noise are negligible in comparison). This front-end noise is assumed to be additive white Gaussian -- an assumption with which experimental results agree. Throughout the study of digital communication systems, the main concern shall be with the problem

of transmitting binary data to the user over an additive Gaussian white noise channel with bit-by-bit signalling (i.e., the space channel). The measure of distortion shall be restricted to information bit error rate P_b. This will yield a measure of information bit reproduction independent of the block structure of a particular code. However, when exact reproduction of blocks of information bits is a necessary system constraint, the codeword error probability P_w may be a more meaningful measure of system performance. In what follows, source characterization is not considered explicitly, but it is assumed that the source has been already efficiently encoded into sequences of binary digits within tolerable distortion. An attempt is then made to maximize the number of these bits which can be sent through the channel to the user unchanged (i.e., minimize the error rate). This decomposes the problem to be considered into two parts. It places the burden of constraining system distortion and maximizing the information rate on the source encoder (e.g., a quantizer, a data compression scheme, etc.) and then choosing a separate channel encoder and decoder to keep the information bit errors over the channel negligible. It can be shown that, in theory, nothing is lost by this decomposition and, therefore, it does not violate the general philosophy of optimizing over the complete system. However, in real systems where practical constraints exist on the source and channel coders, the results may be suboptimal. Nevertheless, it allows one to concentrate on the structure of various channel coding systems without involving the complexity of the overall system design problem initially.

Such a digital communication system is shown in Figure 4. It consists of an information source which puts out binary digits (*information bits*) at a rate r_s bits per second, sequences of which are encoded using a block code of rate R. The output of the encoder is a sequence of binary digits (in general, composed of information bits and parity bits) called *channel bits*, each of which is then assigned by the transmitter one of two analog signals and sent over the channel. At the receiver, each incoming signal is detected as a "one" or a "zero" (bit-by-bit detection with "hard decision") and passed on to a decoder, the outputs of which are the decoded information bits for the user. In a system using bit-by-bit signalling, it is clear that the channel coding is independent of the transmitter/receiver structure and therefore may be optimized separately with no loss. It can be shown that, in this case, the optimum signalling scheme uses antipodal signals (i.e., one analog signal is the negative of the other); these signals are commonly called either a *biorthoginal* or *PSK* signal set and the optimum receiver, a *matched filter*. The output of the filter is a real number, r, with a Gaussian probability distribution (i.e.:

$r = \int_0^{T_C} r(t)s(t)\,dt$, where $s(t)$ is the transmitted signal, $r(t)$

is the received signal corrupted by the channel noise, and T_c is the channel bit duration). A "hard decision" is made by passing this variable, r, through a threshold device which determines its sign. Based on this information, the received signal is detected as a "1" or a "0" (i.e.: $r \underset{\text{"0"}}{\overset{\text{"1"}}{\gtrless}} 0$). Let p equal the probability that a transmitted bit is detected incorrectly at the receiver (by symmetry of the signal set and of the channel noise this probability is independent of whether a "0" or "1" was sent and is therefore well-defined). The probability, p, is, therefore, the channel bit error probability and completely characterizes the channel and the transmitter/receiver structure when "hard decisions" are made at the receiver.

INFORMATION THEORETIC LIMITATIONS ON PERFORMANCE

The digital communications system is represented schematically in Figure 5. Here, the channel and the transmitter/receiver structure have been replaced by a box called a *binary symmetric channel* (BSC). The BSC accepts at its input a channel bit. With probability, p, it delivers to its output the bit inverted and with probability 1-p delivers it unchanged. It can be shown that

$$p = \frac{1}{\sqrt{2\pi}} \int_{\sqrt{\frac{2E_s}{N_o}}}^{\infty} e^{-\frac{x^2}{2}}\, dx = f\left[\frac{E_s}{N_o}\right] = \text{some function of signal-to-noise ratio alone,}$$

where E_s is the energy associated with each channel bit (i.e., the received energy in the signal used to send the bit) and N_o is the average noise power (watts/cycle). Note the simplicity of this result; the channel error rate is simply a function of SNR. This is an extremely pleasant result and is a direct consequence of the white Gaussian channel noise assumption. It should not be expected for general channels other than the space channel. This result allows an explicit relationship to be obtained between information bit error rate and SNR for the BSC.

Assume an average power constraint, S_t, on the transmitter (a design constraint often imposed in practice) and let $S_r = \alpha S_t$ equal the received transmitter power with attenuation α (α depending on antenna size, distance, etc.). Define $E_b \triangleq [nE_s]/k = E_s/R$ as the energy per information bit. If there is no channel coding ($R = k/n = 1$), $E_s/N_o = E_b/N_o$ and $P_b = p = f(E_b/N_o)$ is the information bit error probability. However, in general, with a channel code

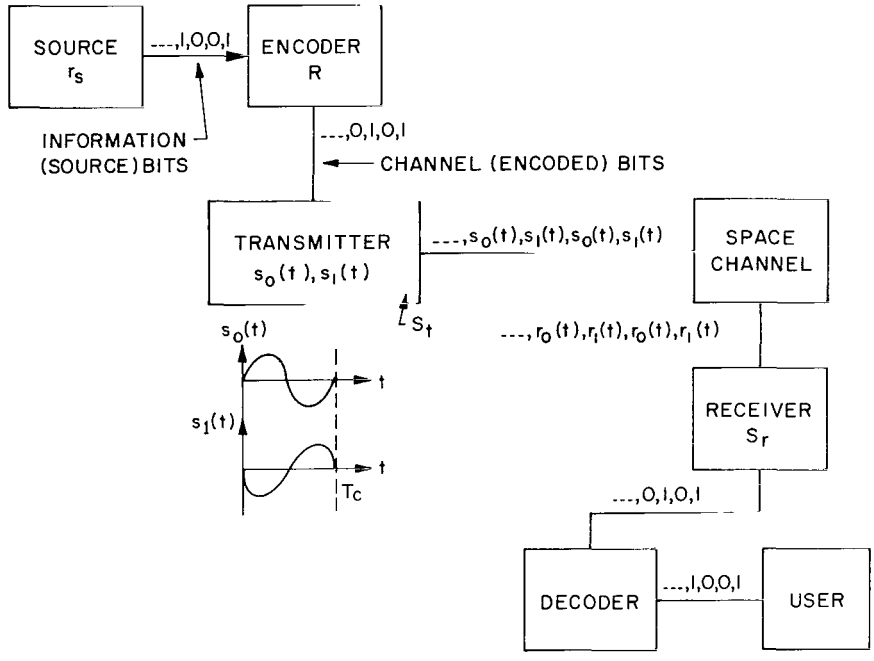

Figure 4.- Digital communication system

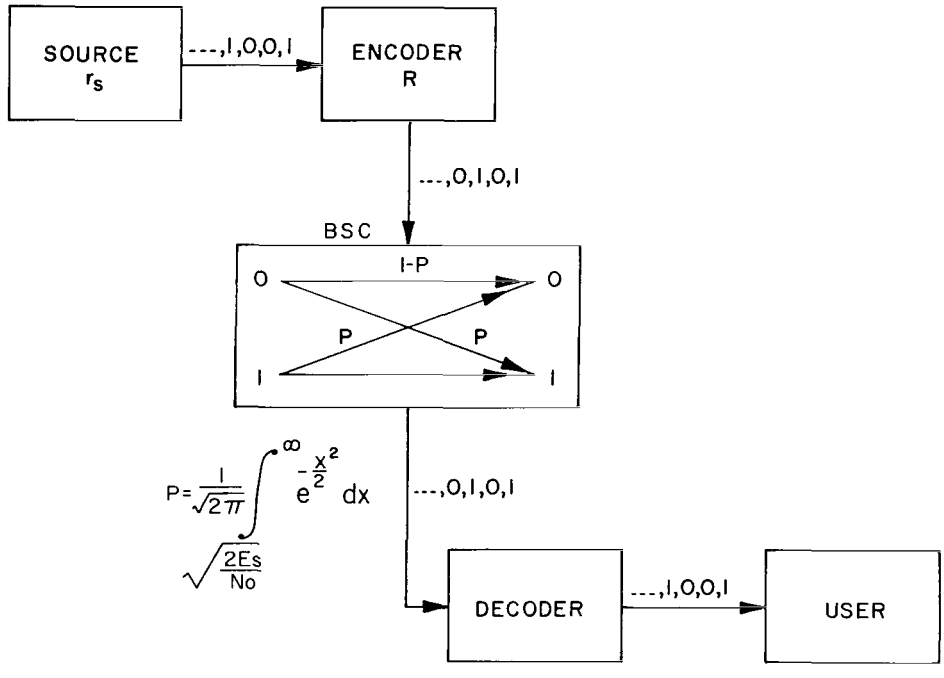

Figure 5.- Digital system schematic

of rate R, $E_s/N_o = R[E_b/N_o]$. In this case, P_b is a function of the channel bit error rate p which, in turn, is a function of E_s/N_o; i.e., $P_b = g(E_s/N_o) = g[R(E_b/N_o)]$. Therefore, for a given code (fixed R), requiring a particular level of performance, P_b implies a value for E_b/N_o. Some symbol manipulation yields $E_b/N_o = nE_s/kN_o = nS_rT_c/kN_o = S_r/N_or = \alpha S_t/N_or$ where $r = k/nT_c$ equals the received information rate in bits per second. From this result, it is clear that one would like to find codes with as small an E_b/N_o as possible, since $E_b/N_o = \alpha S_t/N_or$ implies that a decrease in E_b/N_o would mean less required transmitter power S_t, an increase in the information transmission rate r or allow a smaller α.

Reasonable questions to ask at this point are what is the smallest E_b/N_o that can possibly be hoped for and how difficult is it to achieve? To answer these questions, it is necessary to return briefly to information theory and to the concept of channel capacity. Channel capacity determines the ultimate rate at which one can transmit information over a channel with arbitrarily small error probability. Attempts to transmit reliably above this rate are doomed to failure. For the BSC, the channel capacity is $C_{BSC} = 1 - H(p)$ where H(x) is the binary entropy function (Figure 6a). C_{BSC} as a function of E_s/N_o is sketched in Figure 6b. To determine the minimum E_b/N_o necessary to achieve reliable communication, first fix the code rate at R, choose the smallest E_s/N_o by setting $R = C_{BSC}$ (anything smaller than this would yield $R > C_{BSC}$) and then, using this E_s/N_o, compute $E_b/N_o = E_s/RN_o$. Do this for all R and take the minimum. A graphical interpretation is given in Figure 6c. More formally, set $R = C_{BSC} = 1 - H(f(RE_b/N_o))$. This equation allows the determination of $E_b/N_o|\min,R$ as a function of code rate R. It can be shown that $E_b/N_o|\min,R = h(R)$ is a monotone increasing function of R and that $E_b/N_o|\min = \lim_{R\to 0} h(R) = (2/\pi \ln 2)^{-1} \simeq 0.37$ dB. In comparison, P_b vs. E_b/N_o has been plotted for some simple block codes using bit-by-bit signalling and "hard decision" at the receiver (Figure 7). It can be seen that for small P_b, there is still quite a way to go to achieve $E_b/N_o|\min$ (about 6 or 7 dB). However, it is known from information theoretic arguments that there do exist block codes that operate with negligible error probability and achieve this minimum. As the graph seems to indicate, these codes are probably fairly complex. In many cases, channel encoding is straightforward and relatively simple to implement, even for fairly large block lengths. The complexity generally rests in standard methods of decoding! Therefore, since it is known that good block codes exist, one needs to construct such a class of codes with an efficient decoding algorithm.

An alternate method for determining bounds on the optimum performance of binary coded systems utilizes the results of rate-distortion theory where the average distortion $\bar{D} = P_b$. In this

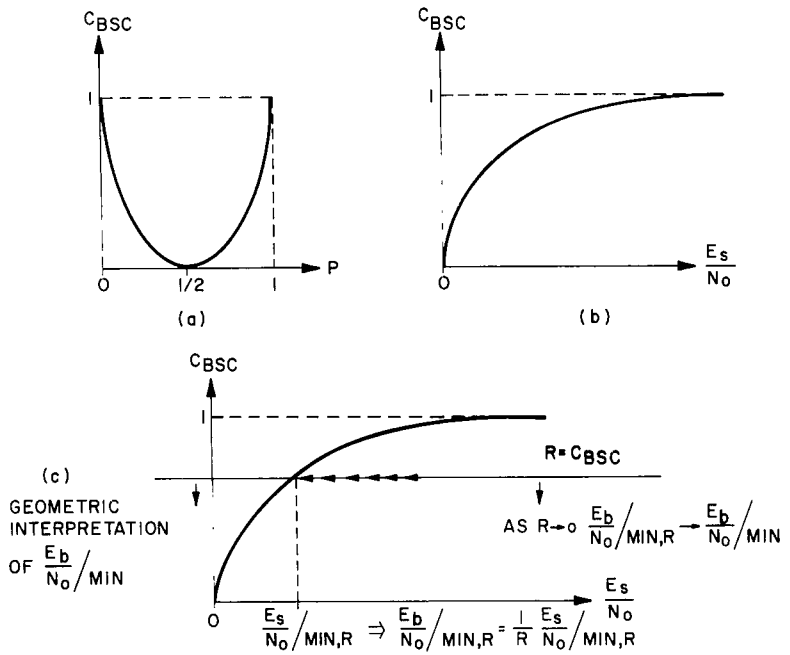

Figure 6.- Channel capacity for BSC

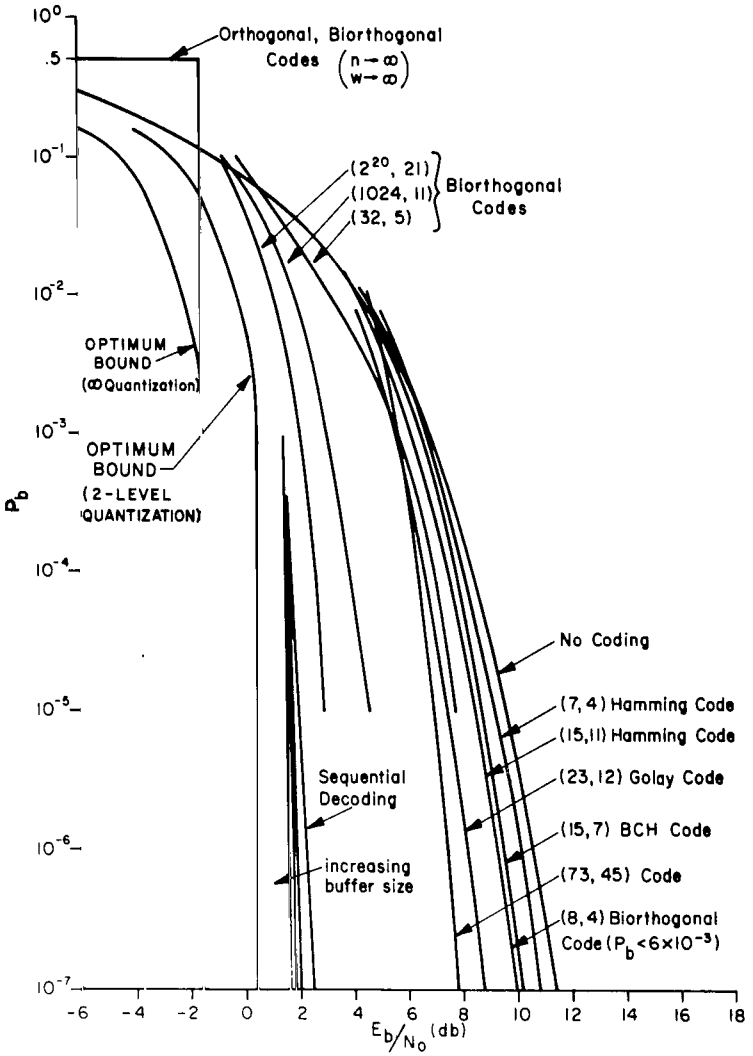

Figure 7. - Graph of P_b vs. E_b/N_o for some codes and comparison with optimum bounds

case, assume a particular transmitter-receiver structure is given. This fixes S_r and T_c which, in turn, determine the channel bit error probability, p. One now asks the following question. What is the maximum information bit rate, r, at which one can transmit over this channel with $P_b \leq \varepsilon$. It can be shown that

$$(*) \qquad r \leq \left. \frac{C_{BSC}(p)}{C_{BSC}(P_b)} \right|_{P_b=\varepsilon} = \frac{1 - H(p)}{1 - H(\varepsilon)} r_c$$

where $r_c = 1/T_c$ is the channel bit rate. In fact, this rate can actually be achieved if coding of arbitrary complexity is allowed. Equation (*) is derived in the Appendix as a direct consequence of rate-distortion theory applied to a binary source. Similar expressions can be derived for channels other than the BSC. Therefore, when constrained to use a particular channel (i.e., fixed p and r_c), equation (*) yields the maximum achievable information bit rate r at which one can transmit information and still maintain $P_b \leq \varepsilon$. Moreover, this result reduces to the first method as a special case if we demand $P_b \to 0$ and maximize r over all possible BSC channels (see the Appendix).

To digress for a moment, an informal derivation of (*) is given which is intuitively appealing. First, as a limiting case, let $P_b \to 0$. Then $1 - H(P_b) \to 1$ and (*) reduces to $r \leq (1-H(p))r_c$ or $r/r_c = k/n = R \leq C_{BSC}(p)$ which, as should be expected, is Shannon's result for error-free transmission. Now in general, Figure 8a illustrates the situation schematically. Given a BSC with crossover probability p, its capacity is $C_{BSC}(p) = 1 - H(p)$ information bits/channel symbol and therefore $(1-H(p))r_c$ is the capacity in information bits/sec. Now if everything within the dotted box is considered as another BSC with crossover probability P_b (Figure 8b), it has capacity $(1-H(P_b))r$ information bits/sec. Since the inner channel forms the "bottleneck" of the communication system, its capacity determines the upper limit on information bit rate, i.e.: $(1-H(P_b))r \leq (1-H(p))r_c$ or $r \leq (1-H(p))/(1-H(P_b))r_c$ "QED".

Up until now in the digital communications system considered, the receiver has been making "hard decisions". As has already been stated, the matched filter output is a continuous variable r which is "quantized" into two levels (positive and negative values) to yield hard decisions at the detector output. Clearly, if some measure of confidence in a "0" or "1" were allowed for, one could make better decisions. One example would be to adopt a binary erasure channel model (see Figure 9). The value of r is detected as one of three levels: a "1", a "0", and an *erasure* X which expresses uncertainty with a value of r too close to the detection threshold (i.e., close to zero). After a complete code word has been received, the information contained in the code structure

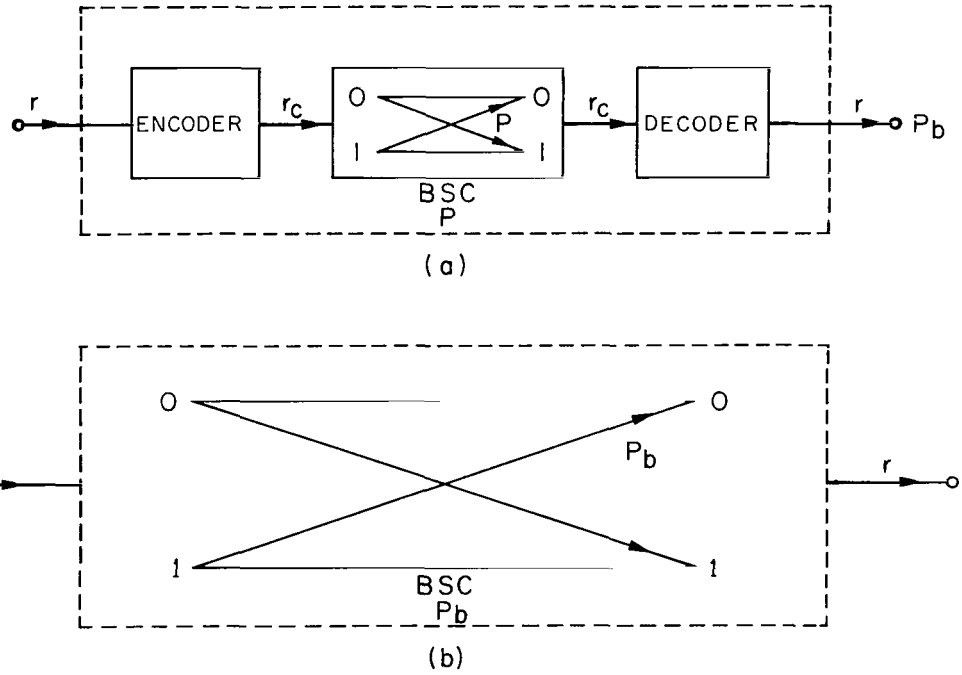

Figure 8.- Schematic interpretation of an informal derivation of (*)

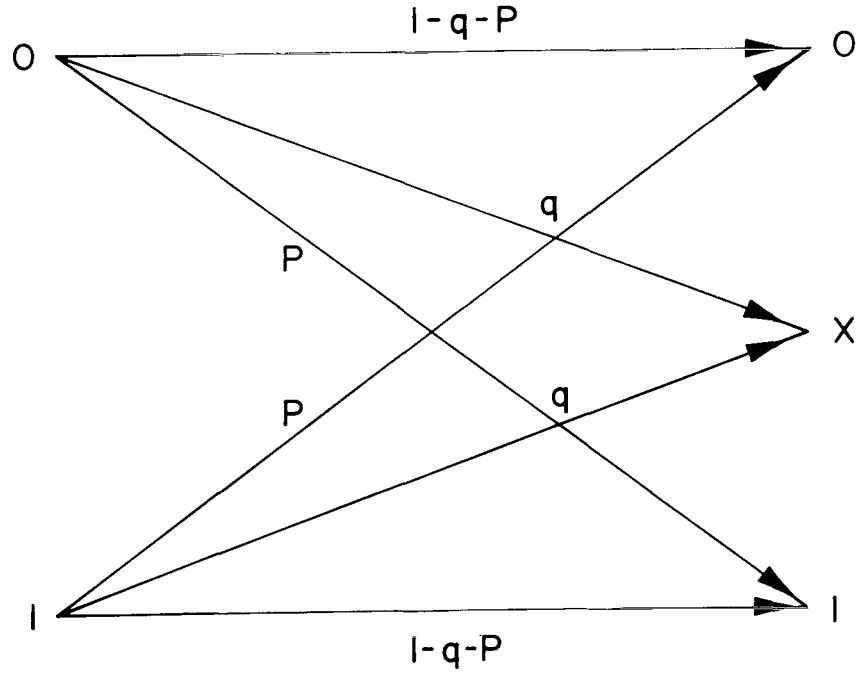

Figure 9.- Binary erasure channel

(e.g., parity checks) and the channel information (erasures) can be combined to help decode the block. In general, r can be quantized to an arbitrary number of levels (each level giving a different amount of confidence in the detected channel bits). How this information is used in decoding depends on the particular code and decoding algorithm used. Certainly the best one can hope to do in extracting information from r will correspond to infinite quantization. By using a simple argument based on the results of information theory -- rate-distortion theory, in particular -- a bound on E_b/N_o as a function of P_b can be derived from an expression similar to (*): $E_b/N_o \geq \ln 2\ [1-H[P_b]]$. This bound is plotted in Figure 7. It can be shown that there exists a code which actually achieves this performance (see the Appendix where this result and some of the other information-theoretic results of this section are proved for the more ambitious reader). Therefore, there is a 2-dB gain in allowing infinite (∞) quantization; one shall see later that, in some cases, eight levels of quantization are enough to achieve this gain. It is interesting to note at this point that Shannon's capacity formula for the continuous channel (where infinite quantization at the transmitter and not just binary signalling is also allowed) yields $E_b/N_o|_{min} = \ln 2$ (obtained for very low code rates R) as in our binary signalling case with infinite quantization just at the receiver. This implies that, theoretically at least, nothing is lost by simply using binary signalling at the transmitter in place of using arbitrarily complex signal sets to represent blocks of the binary data, when willing to use extremely low rate binary block codes (low SNR). It must be pointed out, however, that, in practice, binary signalling might require the use of extremely complex codes and bring added problems (e.g., bit synchronization, etc.) caused by low values of E_s/N_o. With this in mind, the simple transmitter/receiver structure used with binary coding should be explored, complexity measured, and compared with the generally more complex analog modems (modulator/demodulator) available. Note that, although known codes are far from the hard decision optimum bound, this bound might still be approached by these simple codes by using higher-level quantization at the receiver.

ORTHOGONAL CODES

As an example of signal sets more complex than binary signalling, the performance of orthogonal signalling has been plotted in Figure 7. Here, information bits are grouped in blocks of k bits and each block is represented by one of $M=2^k$ orthogonal signals. The corresponding orthogonal signal is transmitted and detected by an optimum receiver (matched filter again). (The use of a complex signal set is somewhat misleading since it can be shown to be equivalent to using a $(2^k,k)$ binary code with binary signalling at the transmitter and ∞-quantization at the receiver.) It has been proven that orthogonal signalling is optimum for the

white Gaussian noise channel with infinite bandwidth (i.e., yields the minimum error rate as $k \to \infty$). The infinite bandwidth constraint is necessary since $k \to \infty$ implies that M = number of orthogonal signals $\to \infty$. It has been shown that the maximum number of orthogonal signals available in time T and bandwidth W is proportional to WT. Since T = ak (i.e., T is proportional to k), then $M=2^k \leq W\,ak$ and therefore $W \geq 2^k/ak \to \infty$. From the graph, it can be seen that the orthogonal codes approach $E_b/N_o|_{min} = \ln 2$ very slowly as k increases (of course, they reach it for $k \to \infty$, but this is not very practical). Also the associated receiver structure is very complex (number of matched filters goes up as 2^k). Therefore, a communication system has been found which is theoretically optimum for the space channel (assuming infinite bandwidth) but hopelessly impractical for large k!

Note that, although orthogonal codes require an infinite bandwidth channel for optimality, Shannon's coding theorem guarantees the existence of binary codes of reasonable rates (e.g., $R \simeq 1/10$) with $P_b \to 0$ for which E_b/N_o is very close to the optimum. Such codes are essentially optimum (in the practical sense) and require only about an order of magnitude more bandwidth than uncoded transmission.

SEQUENTIAL DECODING AND CONVOLUTIONAL CODES

Another example of a binary coded system with bit-by-bit signalling at the transmitter uses convolutional codes with sequential decoding. Convolutional codes *are not* block codes. Information bits are encoded continuously by forming parity checks as sums of information bits from fixed points in the information bit stream (i.e., from a set of taps on a shift register through which the information bit stream passes), whereas, in an (n,k) block code, a block of k information bits at a time are encoded into n channel symbols, and transmitted and the process repeated.

It should be noted that the difference between convolutional codes and block codes is somewhat artificial. Convolutional codes can generally be thought of as a class of very long block codes, since, in practice, the input information stream is periodically terminated for practical reasons such as resynchronization and buffer overflow or excessive computation in decoding.

The resulting codewords of a convolutional code can be represented by a binary tree structure. The sequential decoding procedure amounts to a tree-searching algorithm to find the codeword which was most likely sent. The performance of such a scheme is fairly impressive. For code rates less than some prescribed quantity R_{comp} (a number which depends on the channel), one can achieve the negligible error probability guaranteed by the coding theorem. The price paid is an amount of computation and a decoding buffer size which is large (*but not exponential*). This is a

good example of the trade-off between a simple encoder and a complex decoding algorithm. However, simple calculation shows that with 2-level quantization at the receiver, R_{comp} constrains $E_b/N_o|min = 2[(\pi \ln 2/2)]$, a 3-dB departure from the general theoretical limit. Even if one allows ∞-level quantization at the receiver, there is still a 3-dB loss in performance; i.e., $E_b/N_o|min = 2 \ln 2$. Some typical performance curves for sequential decoding have been sketched onto the graph of Figure 7. In many practical situations, eight levels are all that is usually required to come close to the performance of ∞-level quantization. The reason for the family of curves is the large random variation of computation required by the decoding algorithm. In fact, the theoretical scheme allows unrestricted (impractical) decoder buffer size and amount of computation. The deviation from the theoretical performance $E_b/N_o|min = 2 \ln 2$ is, for the most part, due to the practical constraints on these quantities.

The large gap in performance between convolutional codes with sequential decoding and block codes of short length plotted in Figure 7 is somewhat misleading. There is nothing magical about convolutional codes: they are essentially just a class of big block codes. They have some nice properties (e.g., a tree structure) which are profitably exploited by sequential decoding. There are other known classes of block codes which have been constructed with nice structures which appear to be good. Work is being done in trying to develop efficient decoding algorithms for these codes.

For one such class of codes, BCH codes, a fairly efficient *suboptimum* (algebraic) decoder exists. Using bit-by-bit signalling with 2-level quantization at the receiver, simulation has shown that the (255, 123) BCH code's performance is only about 2 dB below sequential decoding with 2-level quantization (and about 2 dB above the best short-length block code plotted in Figure 7). However, when compared with sequential decoding with ∞-level quantization (achieved in practice with about 8-levels) there is a 5-dB gap. The sequential decoding algorithm can easily incorporate the additional information gained by multi-level quantization, whereas this BCH decoder cannot. Also, the average amount of computation for BCH decoding is larger than for sequential decoding, but does not exhibit the large random variation of the latter. Note that this is an example of just one type of possible BCH decoder.

Sequential decoding is an adaptive search technique. It is essentially non-algebraic in nature, exploiting the probabilistic information at the receiver. It is not designed for the worst-case channel noise, but instead adapts to changing channel disturbance. Schemes of this general type could quite possibly be used to decode efficiently some of these other classes of "good" block codes (possibly similarly constrained to operate at rates a fraction

of capacity). It would appear that the most efficient algorithm would, in general, exploit both the underlying code structure (if any) and the probabilistic information obtained from the receiver in the decoding.

Just how much of the performance of convolutional coding with sequential decoding is fundamental and what portion is just due to the relatively small amount of experimental effort expended to find efficient decoding algorithms of this general type for other classes of codes remains to be seen.

CONCLUSIONS

Fundamental to the communication system design problem is efficient source characterization. Having modelled the source, as much redundancy was removed from source outputs as was possible thereby "compressing" the data. Faced with a noisy transmission medium (in this case, space with front-end receiver noise), the task was to transmit this "compressed" source reliably to the user. At the start, both analog and digital communication systems were considered. When the general problem was recognized as being unwieldy, the decision was made to pursue only one area -- digital communication -- since it is less well-known to the communication system designer than analog techniques.

While realizing that, in general, the results would be suboptimum for real systems, it was assumed that the source had been efficiently encoded into binary digits within tolerable distortion and transmitted to the user over a noiseless channel. With this assumption, the space channel was approximated by a noiseless channel using channel coding to "clean it up". This meant making the information bit error rate on the channel negligibly small. After some simple calculations, it was seen that for a fixed code rate, R, the information bit error rate was a function of the information bit energy and that there was a trade-off between this energy, transmitter power, channel attenuation, and the received information bit rate. The simplicity of the relations derived relating error rates to SNR is due to the form of the statistics for the space channel. They allowed simple bounds on optimum system performance to be obtained. Such results should not be expected on more general channels, such as bursty channels, fading channels, or other HF channels. Also, systems designed to function efficiently for a space channel should not be expected to work well on other types of channels.

Implicit in the discussion thus far has been a trade-off between code rate, R, and the actual information rate, r, at the receiver. It has been seen from theoretical considerations that maximum information rate is achieved with extremely low code rates (i.e., $R \to 0$). What does all this mean heuristically? Suppose

that based on subjective considerations, a suitable error rate P_b is decided on (e.g., $P_b \simeq 10^{-4}$). With no channel coding (R=1), this implies a particular $E_s|_{u.c.} = E_b|_{u.c.}$ which yields $r_{u.c.} = S_r/E_b|_{u.c.}$. Now, consider using channel coding which yields $r_c = S_r/E_b|_c$. Therefore $r_c > r_{u.c.}$, if and only if, $E_b|_c < E_b|_{u.c.}$. In general, given two codes, 1 and 2, then $r_2 > r_1$ if and only if $E_{b2} < E_{b1}$. Since $E_{s1} = R_1E_{b1}$ and $E_{s2} = R_2E_{b2}$, we have $E_{s2}/R_2 < E_{s1}/R_1$. Therefore, higher information rate, r, means lower information bit energy E_b which since $E_s = S_rT_b = RE_b$, means lower channel bit energy. This last relation implies increased channel bit error rate p (lower SNR on the channel due to the decrease in channel bit duration T_c means less allowed energy per channel bit). The trade-off is clear. The more code redundancy 1-R added to combat channel noise, the smaller E_s becomes and channel errors become more likely. Therefore, the amount of noise immunity obtained from the code must more than offset the accompanying increase in channel errors caused by the decrease in channel bit energy. This form of code performance loss is called *code rate loss* and must be compensated for by a decrease in information bit error rate. This then can be turned into increased information rate, r, by lowering the required signal energy.

Earlier in this section, two quantities for information rate, r_s and r, were defined. The quantity r_s is the source rate and is used to characterize the source. It is a property of the source and the source encoder and is independent of the remainder of the communication system. The quantity r is the information bit rate achieved at the receiver and depends on the received transmitter power and information bit energy. One would like $r = r_s$. In general, however, $r < r_s$ since most interesting sources have a very high data rate; in this case, a data buffer is assumed or data is thrown away. If $r > r_s$, the channel rate is higher than necessary and a simpler system could probably be used.

An advantage of block coding schemes over uncoded systems not previously mentioned is their immunity to imprecise channel model statistics (especially on the tails of the distribution where the Gaussian form given by central limit theorems tends to break down.) In calculating error performance for block codes, one considers the behavior of long sequences of channel symbols and, therefore, is treating "averaged" or "smoothed" quantities as one does with laws of large numbers. Such operations depend on robust statistical properties of sequences (and sums) of random variables (the channel symbols) and are therefore essentially independent of individual channel symbol statistics.

As was pointed out earlier, the encoding of long block codes can be implemented fairly easily. Ease in decoding is generally the problem and, consequently, a large amount of research effort

is directed toward this end. Of course, new classes of codes are continually being searched for with the hope that they might have simple decoding procedures. It should be mentioned, however, that depending on the particular system constraints, fairly complex decoding algorithms may be tolerable. For example, in space communications where spacecraft transmitter power and space are at a premium, a simple encoder might be used to transmit data at high rates to ground stations. Data could then be either stored on tape or processed real-time by computer, depending on the noise level, by using a complex decoding algorithm.

In this report, block codes have been considered which allow one to correct channel errors in the received information (error-correcting codes). Another alternative would be to use simpler block codes which simply detect channel errors and request the retransmission of the information in error. Under such circumstances, the quantities of interest would include the average transmission rate and the probabilities of detected and undetected errors. Such a scheme is just a special case of a communication system with a feedback link (which may be noisy or noiseless). How does such a scheme compare with the schemes thus far considered? This is a legitimate question which should be considered (e.g., telephone lines are natural 2-way links).

Finally, as has already implied, coding is not only relevant for deep space probes where received power is weak and SNR is low. Depending, of course, on the particular system design constraints, coding can be used when data rates are so high that extremely low E_b is available; remember $r = S_r/E_b$ bits/sec. Of course, before starting to use channel coding here, it is assumed that the data source itself has been efficiently encoded so that r_s is not high due to remaining redundancy in the data. Once this has been checked and the receiver gain has been turned up all the way and still $r << r_s$, one must turn to channel coding if a higher received data rate is desired.

APPENDIX

SUMMARY

When the results of rate-distortion theory (ref. 1), are used, an upper bound on information rate (bits/sec) as a function of bit error probability can be calculated. A novel interpretation of the trade-offs between bit error rate, signal-to-noise ratio (SNR), and information rate is presented which is particularly appealing to the communication system designer.

Rate-distortion theory extends Shannon's results for error-free transmission (ref. 2) to include bounds on the performance of communication systems which tolerate some distortion in reproduction to improve data rate or reduce system complexity. A cursory treatment of the theory relevant to the problem at hand is presented.

RATE-DISTORTION THEORY

It is desired to transmit over a communication channel to a user the outputs from an information source. As a measure of the fidelity of user reproduction, a single-letter distortion function d(u,v) is defined which measures the relative unhappiness of receiving letter v when u was actually sent. The channel is assumed to be memoryless and is therefore completely characterized by a set of transition probabilities Pr[v/u] = Pr[v received/u sent] for all u,v and specifies the entire communication system between the information source and the user. It is assumed that sequences of k source letters are encoded into a block code of n channel symbols before being sent over the channel. Define

$$D(\underline{u},\underline{v}) = \frac{1}{K} \sum_{i=1}^{k} d(u_i, v_i)$$

where the subscripts denote successive source letters in a particular code block and the average distortion per letter

$$D = \sum_{\underline{u},\underline{v}} D(\underline{u},\underline{v}) \Pr[\underline{u},\underline{v}]$$

where the sum is taken over all possible source and received sequences of length k. The rate-distortion function R(D*) is defined as the minimum mutual information between the source and the user (*calculated for a single-channel use*) minimized over all channels subject to the constraint that the average source letter distortion

$$\sum_{u,v} d(u,v)\ \Pr[v/u]\ \Pr[u] \leq D^*$$

where Pr[v/u] are channel transition probabilities and Pr[u] the source distribution. It can be shown (ref. 1) that given a channel of capacity C with tolerable distortion level D* (i.e., $D \leq D^*$), the information transmission rate k/n (information symbols/channel symbol) is upper-bounded by C/R(D*). In fact, it is possible to approach this rate arbitrarily closely with suitably complex encoding of the source letters (i.e., large n). The above results allow us to interpret R(D*) as the *equivalent rate of the source*.

DERIVATION OF THE BOUND

If one lets the source be the independent binary source with equiprobable letters and the single-letter distortion $d(u=i,v=j) = 1-\delta_{ij}$, then

$$D = \frac{1}{K} \sum_{i=1}^{k} P_{bi} = P_b$$

where P_{bi} equals the probability of bit error in the i'th position of the encoded source sequence and, therefore, P_b is the average bit error probability. For this source, it has been shown (ref.1) that $R(D) = R(P_b) = 1 - H(P_b)$ where H(x) is the binary entropy function $H(x) = -x \log x - (1-x) \log (1-x)$. Therefore we have:

$$\frac{k}{n} \leq \frac{C}{R(P_b)} = \frac{C}{1-H(P_b)} \qquad \text{(1a)}$$

where equality can be approached arbitrarily closely by increasing n (i.e., long block codes). Now if one lets T_c equal the channel bit duration, he has:

$$\frac{k}{nT_c} \leq \frac{C}{1 - H(P_b)} \quad \frac{1}{T_c}$$

or

$$r \leq \frac{C}{1 - H(P_b)} r_c \tag{1b}$$

where r = information bits per second, and r_c = channel bits per second.

A heuristic derivation of Eq. (1) has already been given in the main body of the paper.

If, as a communication system designer, a particular modem (transmitter-receiver structure) has been given, then the channel capacity C, the channel bit duration T_c, and, in turn, r_c, are fixed in the design. In this case, Eq. (1) yields the maximum information bit rate at which one can transmit over this channel, no matter how cleverly the source is encoded, while maintaining a bit error rate P_b. However, if one has some freedom in designing the modem, the maximum information bit rate can be derived from Eq. (1), if one maximizes the right-hand side over all allowable modems (e.g., all modems with an average transmitter power constraint S_t):

$$(1^*) \qquad r \leq \max_{\substack{[\text{allowable}] \\ \text{modems}}} \left(\frac{C}{1 - H(P_b)} r_c \right)$$

which implies the maximization problem:

$$\max_{\substack{[\text{allowable}] \\ \text{modems}}} \frac{C}{T_c}$$

APPLICATION TO THE SPACE CHANNEL

Some important examples will now be treated.

Case I — Space Channel/Two-Level Quantization

The encoded bit sequences from the source are transmitted using bit-by-bit signalling (optimum binary antipodal signalling) over a white Gaussian additive noise channel with average noise power N_o (single-sided) and average transmitted signal power S_t. The received power $S_r = \alpha S_t$ where α is a function of distance from the transmitter, antenna gain, etc. At the receiver, the signal is quantized to one of two levels and a "hard decision"

made as to whether a "0" or a "1" was sent. For this case, with fixed S_t and α, one has:

$$\max_{\substack{[\text{allowable}] \\ \text{modems}}} \frac{C}{T_c} = \max_{\substack{[\text{BSC modems}] \\ \text{fixed } S_r}} \left(\frac{C_{BSC}(p)}{T_c} \right)$$

where $C_{BSC}(p) = 1 - H(p) = 1 + p \log p + (1-p) \log (1-p)$

and $p = \frac{1}{\sqrt{2\pi}} \int_{\sqrt{\frac{2S_r T_c}{N_o}}}^{\infty} \exp\left[-\frac{X^2}{2} \right] dx$ = function of S_r, T_c.

It can be shown that $g(S_r,T_c) = \frac{C_{BSC}(S_r,T_c)}{T_c}$ is a monotone-decreasing function of T_c. Therefore,

$$\max_{\substack{[\text{BSC modems}] \\ \text{fixed } S_r}} \left(\frac{C_{BSC}(S_r,T_c)}{T_c} \right) = \lim_{T_c \to 0} \left(\frac{C_{BSC}(S_r,T_c)}{T_c} \right) = \left. \frac{dC_{BSC}}{dT_c} \right|_{T_c = 0}$$

$$= \frac{S_r/N_o}{\frac{\pi}{2} \ln 2}$$

Inserting in Eq. (1*):

$$r \leq \frac{\alpha S_t}{N_o} \left(\frac{1}{\frac{\pi}{2} \ln 2 (1-H(P_b))} \right)$$

where equality can be approached arbitrarily closely by complex encoding (long block codes).

Case II — Space Channel/Unquantized Receiver

Same as Case I with ∞-level quantization at the receiver. It is easily shown that the capacity of the unquantized channel $C_{U.Q.}$ is:

$$C_{U.Q.} = \frac{1}{\sqrt{\pi N_o}} \int_{-\infty}^{+\infty} \left[\exp \frac{(x-\sqrt{E_s})^2}{N_o} \right] \log_2 \left[\frac{2}{1 + \exp\left[-4x \frac{\sqrt{E_s}}{N_o}\right]} \right] dx$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \exp\left[\frac{-x^2}{2}\right] \log_2 \left[\frac{2}{1 + \exp\left\{-4\left[\sqrt{\frac{E_s}{2N_o}}\, x + \frac{E_s}{N_o}\right]\right\}} \right] dx$$

$$= -E_G \left\{ \log_2 \left[\frac{1 + \exp\left\{-4\left[\sqrt{\frac{E_s}{2N_o}}\, x + \frac{E_s}{N_o}\right]\right\}}{2} \right] \right\}$$

where E_G is the expectation using the Gaussian probability distribution with zero mean and unit variance and $E_s = S_r T_c$. Once again, it is easy to show that $C_{U.Q.}/T_c$ is monotone-decreasing.

Therefore,

$$r \leq \frac{\lim\limits_{T_c \to 0} \frac{C_{U.Q.}}{T_c}}{1 - H(P_b)} = \frac{\left.\frac{dC_{U.Q.}}{dT_c}\right|_{T_c = 0}}{1 - H(P_b)} = \frac{\alpha S_t}{N_o} \left(\frac{1}{\ln 2\, [1-H(P_b)]} \right)$$

Since

$$\frac{E_b}{N_o} = \frac{nE_s}{kN_o} = \frac{nS_r T_c}{kN_o} = \frac{S_r/N_o}{r} ,$$

where E_b and E_s are information bit and channel bit energy, respectively, one can reinterpret the preceding results as an optimum bound on P_b versus E_b/N_o.

Case I

$$\frac{E_b}{N_o} \geq \frac{\pi}{2} \ln 2\, [1-H(P_b)]$$

or

$$P_b \geq H^{-1}\left(1 - \frac{E_b/N_o}{\frac{\pi}{2}\ln 2}\right).$$

Case II

$$\frac{E_b}{N_o} \geq \ln 2\ [1-H(P_b)] \text{ or } P_b \geq H^{-1}\left(1 - \frac{E_b/N_o}{\ln 2}\right).$$

Therefore, no matter how cleverly one encodes the information source, one can only achieve a $10 \log_2 \pi/2 \simeq 2$ dB decrease in required E_b/N_o by going from 2-level to ∞-level quantization at the receiver. Of course, in practical situations, one might very well consider trading some coding complexity for an increased number of quantization levels (in many cases no more than 8 levels of quantization are necessary to approach the performance of the unquantized receiver). Figure 7 is a graph of P_b versus E_b/N_o for some typical codes plotted for comparison with these optimum bounds. The orthogonal and biorthogonal codes use ∞-level quantization (matched filtering) at the receiver as do the convolutional codes decoded by sequential decoding. All the other codes have been plotted assuming 2-level quantization. One can see that for small P_b, one still has quite a way to go (about 6 or 7 dB) to achieve the minimum E_b/N_o guaranteed by the bound when restricted to use one of the given (n,k) block codes of moderate size. As new codes and decoding techniques are found, the graph of Figure 7 can be used to achieve a partial ordering of the codes according to required E_b/N_o (or, equivalently, information rate) for a given error rate P_b. Of course, complexity of a particular coder and decoder implementation and the amount of computation per information symbol must also be considered.

Many communication system designers feel that they have been short-changed by the communication theorist in that he has not stated his results in terms meaningful to the practicing engineer. For example, the practicing system design engineer is not directly concerned with the particular mapping of source symbols to codewords used to match the information source to the channel or in the code information rate $R = k/n$. In many cases, he will tolerate any code (within the system constraints of complexity, computation time, etc.) and is really interested in overall system performance in terms meaningful to him -- such as achievable data rate and error rate. In a given design situation, he may or may not have the freedom to choose all parts of the system. In general, however, he is seeking to optimize the overall system over a class of codes *and* allowable modems. This last degree of freedom may be interpreted as a variable channel capacity, C, subject to some physical constraint (e.g., average transmitter power).

As examples of this general approach, two cases were given. The technique can be applied to other channels of interest in a straightforward way. The simplicity of the results for the two cases given is due primarily to the additive Gaussian white noise assumption. Also, the results can be extended to other sources and fidelity criteria (although the calculations, in general, would be more difficult). For a good introductory treatment of rate-distortion theory and methods used to calculate rate-distortion functions and appropriate bounds for different distortion criteria, see Gallager (ref. 3).

REFERENCES

1. Shannon, C. E.: Coding Theorems for Discrete Source with a Fidelity Criterion. IRE Nat. Conv. Record, part 4, 1959, pp. 142-163.

2. Shannon, C. E. and Weaver, W.: The Mathematical Theory of Communication. Urbana, Illinois, University of Illinois Press, 1964.

3. Gallager, R. G.: Information Theory and Reliable Communication. New York: Wiley, 1968, pp. 442-502.

"The aeronautical and space activities of the United States shall be conducted so as to contribute . . . to the expansion of human knowledge of phenomena in the atmosphere and space. The Administration shall provide for the widest practicable and appropriate dissemination of information concerning its activities and the results thereof."

—NATIONAL AERONAUTICS AND SPACE ACT OF 1958

NASA SCIENTIFIC AND TECHNICAL PUBLICATIONS

TECHNICAL REPORTS: Scientific and technical information considered important, complete, and a lasting contribution to existing knowledge.

TECHNICAL NOTES: Information less broad in scope but nevertheless of importance as a contribution to existing knowledge.

TECHNICAL MEMORANDUMS: Information receiving limited distribution because of preliminary data, security classification, or other reasons.

CONTRACTOR REPORTS: Scientific and technical information generated under a NASA contract or grant and considered an important contribution to existing knowledge.

TECHNICAL TRANSLATIONS: Information published in a foreign language considered to merit NASA distribution in English.

SPECIAL PUBLICATIONS: Information derived from or of value to NASA activities. Publications include conference proceedings, monographs, data compilations, handbooks, sourcebooks, and special bibliographies.

TECHNOLOGY UTILIZATION PUBLICATIONS: Information on technology used by NASA that may be of particular interest in commercial and other non-aerospace applications. Publications include Tech Briefs, Technology Utilization Reports and Notes, and Technology Surveys.

Details on the availability of these publications may be obtained from:

SCIENTIFIC AND TECHNICAL INFORMATION DIVISION

NATIONAL AERONAUTICS AND SPACE ADMINISTRATION

Washington, D.C. 20546